

From Black Box to Trustworthy AI: A Secure Framework for Explainable Cybersecurity Decision-Making

Syedmostafa Safavi

School of Technology

*Asia Pacific University of Technology
and Innovation, Kuala Lumpur, Malaysia*

Email: seyed.safavi@apu.edu.my

Mohamed Shabbir Hamza Abdulnabi

School of Technology

*Asia Pacific University of Technology
and Innovation, Kuala Lumpur, Malaysia*

Email: mohamed.shabbir@apu.edu.my

Muhammad Ehsan Rana

School of Computing

*Asia Pacific University of Technology
and Innovation, Kuala Lumpur, Malaysia*

Email: muhd_ehsanrana@apu.edu.my

Shahab Alizadeh

School of Technology

*Asia Pacific University of Technology
and Innovation, Kuala Lumpur, Malaysia*

Email: shahab.alizadeh@apu.edu.my

Abstract—The escalating amalgamation of Artificial Intelligence (AI) within the domain of cybersecurity has engendered notable progressions in the realms of threat identification and mitigation. Nevertheless, the widespread deployment of opaque black-box AI models, which inherently lack clarity regarding their decision-making methodologies, poses substantial difficulties for cybersecurity practitioners. This scholarly article seeks to address this particular constraint by introducing an innovative secure framework aimed at elucidating AI processes within the cybersecurity context. The proposed framework amalgamates various explainability techniques, including SHAP, LIME, and Grad-CAM, to elucidate the rationale underlying AI decision-making. Additionally, it integrates comprehensive security layers that encompass data safeguarding, adversarial mitigation, and model provenance to uphold the integrity and robustness of AI systems. Moreover, trust-enhancing mechanisms, which include thorough logging, human-in-the-loop supervision, and model certification, are incorporated to cultivate confidence in AI-driven cybersecurity determinations. This framework aspires to augment transparency and reliability in essential security applications such as malware identification, phishing remediation, and AI-facilitated Security Operations Centers (SOCs), thereby paving the path for a more credible and efficacious application of AI within the cybersecurity paradigm.

Index Terms—Explainable AI, Trustworthy AI, Secure Machine Learning, Cybersecurity Decision Support, AI in Cybersecurity, Adversarial Defense, Model Provenance.

I. INTRODUCTION

The digital landscape is currently experiencing an unprecedented escalation in both the frequency and complexity of cyber threats, which significantly affect a multitude of sectors, ranging from financial services to healthcare systems [1]. Advanced threat actors utilize highly sophisticated methodologies, thereby necessitating the establishment of resilient cybersecurity frameworks that are capable of swift

and precise detection and response [2]. Intrusion Detection Systems (IDS) have evolved into key features of cybersecurity landscapes, utilizing artificial intelligence (AI) and machine learning (ML) systems to explore network traffic and user engagements for swift threat discovery. While these systems enhanced by artificial intelligence frequently attain commendable precision in detection, they generally function as opaque mechanisms, offering inadequate clarification regarding their decision-making frameworks [3], [4]. This lack of transparency poses significant obstacles for security analysts, who require an in-depth understanding of the rationales behind each detection to validate, refine, and enhance model outputs, especially in sensitive operational environments. The deficiency in interpretability concerning artificial intelligence systems can lead to a multitude of complications that may manifest as an increased incidence of false positives, along with significant obstacles in effectively addressing complex threats, which arise primarily from the inherently ambiguous nature of the reasoning processes employed by these models. This pervasive uncertainty has the potential to significantly undermine the level of trust that users place in the system, ultimately hindering the efficient validation and enhancement of models, while also potentially obscuring critical vulnerabilities that may exist within the framework of the system itself. Moreover, ethical quandaries, which include biases that may be deeply entrenched within these opaque models, further exacerbate the challenges associated with their interpretation and application [5].

In the realm of cybersecurity, the attributes of transparency, security, and reliability emerge as indispensable for the proficient and ethical deployment of artificial intelligence technologies. The aspect of interpretability empowers security analysts, granting them the essential ability to comprehend the

*Corresponding author: seyed.safavi@apu.edu.my

underlying justifications that inform AI-generated decisions, which in turn facilitates their capacity to critically evaluate the accuracy of alerts and systematically improve the models over time through iterative processes of learning and adaptation [6]. Furthermore, transparency plays a crucial role in fostering trust among security personnel and various stakeholders involved in the cybersecurity landscape, thereby promoting accountability while simultaneously aiding in the adherence to increasingly rigorous regulatory standards that govern the use of such technologies [7]. Additionally, grasping the elaborate decision-making processes in artificial intelligence systems can substantially boost security measures by assisting in revealing potential vulnerabilities, thus reinforcing defenses against a wide array of adversarial challenges that might attempt to exploit these shortcomings. Considering the challenges imposed by opaque AI systems in cybersecurity, there is an urgent demand for a framework that inherently integrates both explainability and security. Such a framework would equip security practitioners with the essential insights and authority necessary to leverage AI's capabilities in a manner that is both reliable and effective, ultimately resulting in more robust and secure cyber defenses [6].

II. PROPOSED FRAMEWORK

To alleviate the problems related to inscrutable AI in cybersecurity, we support the formation of a comprehensive system that incorporates clarity, safety measures, and trust components. The theoretical architecture of this framework encompasses several pivotal components functioning synergistically (refer to Figure 1).

One way to evaluate AI functionality involves an assortment of methods, particularly emphasizing SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), and Grad-CAM (Gradient-weighted Class Activation Mapping). SHAP affords both global and local assessments of feature significance, thereby enabling security analysts to comprehend the contribution of various features (e.g., characteristics of network traffic, headers of emails) to the predictions rendered by the AI. This understanding is essential for deriving insights into the rationale behind the classification of a specific activity as either malicious or benign. LIME provides localized explanations by simulating the behavior of a complex model with a more straightforward, interpretable model tailored for individual instances. This approach proves particularly advantageous for auditing discrete cybersecurity determinations, such as elucidating the reasons a specific email was categorized as phishing. Grad-CAM generates visual explanations through the production of heatmaps that denote critical regions within the input data that influenced the AI's decision-making process. As an illustration, in terms of network intrusion detection, Grad-CAM could underscore specific characteristics in network traffic that initiated an alert. These methods of explainability can be universally applied across diverse cybersecurity scenarios, encompassing anomaly detection in network traffic, malware classification based on file attributes, and the identification of

phishing attempts predicated on the content and metadata of emails.

The framework integrates multiple security layers to guarantee the robustness and integrity of the AI system. Mechanisms for safeguarding data, encompassing the encryption of sensitive training and operational datasets, alongside anonymization methodologies, are implemented to protect against data breaches and violations of privacy. Approaches utilized for bolstering adversarial resilience, such as adversarial training that familiarizes the model with deliberately constructed harmful inputs, in conjunction with input sanitization aimed at eradicating potentially detrimental data, are integrated to strengthen the model's defenses against adversarial threats. In addition, it is vital to perform an exhaustive and organized documentation of where the model comes from, requiring careful logging of a broad spectrum of information about the model's roots, the specific datasets that have been applied during the training stages, and the sophisticated and often intricate approaches that have been adopted in the model's formation. This thorough documentation method is critically important because it contributes fundamentally to preserving the model's integrity and authenticity, while concurrently enabling the detection and analysis of any potential vulnerabilities or biases that could develop during the model's functional duration.

Trust mechanisms are undeniably critical components within the framework that has been proposed for this particular study. An exhaustive logging protocol that captures all actions carried out by the AI system is indispensable; this protocol encompasses a detailed record of inputs, outputs, and the rationale behind various decisions made by the system, thereby significantly enhancing the auditability of operations and enabling comprehensive investigations into any incidents that may arise. Furthermore, processes that incorporate human oversight, commonly referred to as Human-in-the-Loop (HITL) methodologies, have been established to guarantee that human validation is applied to critical determinations made by the AI. This strategic approach serves to mitigate the inherent risks that are associated with an overreliance on AI technologies and empowers security analysts, allowing them to utilize their expertise effectively in complex or uncertain scenarios that may demand nuanced human judgment. In addition, the concept of model certification is introduced, wherein the AI model is subjected to rigorous testing and evaluation against established benchmarks that assess its security, explainability, and reliability. Such a certification process can significantly foster and cultivate confidence among various stakeholders regarding the credibility and trustworthiness of the model.

This secure-XAI framework demonstrates substantial potential for practical application across a multitude of domains within the field of cybersecurity, showcasing its versatility and adaptability. Specifically, in the realm of malware detection, the framework proves to be highly effective in elucidating the rationale behind the classification of a particular file as malicious, thereby highlighting specific segments of code or behavioral patterns that triggered the alert mechanism. In addition, in the sphere of responding to phishing dangers,

this framework can clarify the signs that were instrumental in detecting a phishing email, including suspicious links or abnormal sender details that do not conform to what is expected. In addition, the utilization of this model in AI-augmented Security Operations Centers (SOCs) can substantially boost the effectiveness of security analysts by offering them extensive, contextually pertinent clarifications for the alerts they receive, thereby nurturing more insightful decision-making processes and allowing for a faster and more impactful response to new threats within the cybersecurity arena.

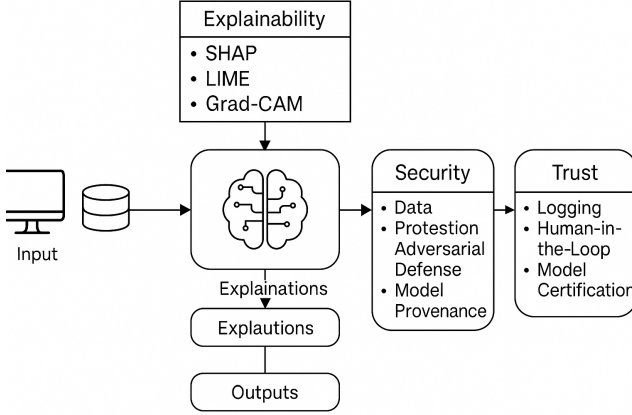


Figure 1. Conceptual architecture of the proposed Secure-XAI framework for cybersecurity

III. RELATED WORK

Safavi and Shukur (2020) presented "CenterYou: A Cloud-Based Methodology for the Simplification of Android Privacy Management." CenterYou employs a cloud-centric architectural framework, utilizing pseudo data techniques to protect user privacy on Android devices through the interception of permission requests and the provision of fabricated data to applications [8]. This methodology contrasts sharply with the potentiality of our proposed framework, which is meticulously designed to operate within edge-centric or decentralized secure-AI paradigms, thereby enhancing cybersecurity. While CenterYou is contingent upon cloud processing for privacy-related determinations, our framework envisions the capability to execute analysis and decision-making in proximity to the data source or through a distributed framework for the detection of cybersecurity threats, thereby potentially yielding advantages in terms of diminished latency and improved data locality. The reliance of CenterYou on cloud computing diverges from the benefits presented by edge or decentralized secure AI in the realm of cybersecurity, where processing at the network edge can facilitate quicker response times and minimize overhead.

In "Investigating the Role of Blockchain in Enhancing Home Security: A Comprehensive Study" (2023), Safavi et al. examined the application of blockchain technology in securing Internet of Things (IoT) devices and managing ownership within smart home ecosystems [9]. Their research underscores the trust-enhancing attributes of blockchain, including immutability and transparency, within decentralized environments. These properties are relevant to the trust mechanisms

in our proposed framework, particularly in decentralized cybersecurity scenarios where blockchain could potentially be used for secure logging or decentralized model verification.

Studies focusing on Geographic Information Systems (GIS) or radiological studies, while utilizing AI, address distinct challenges related to spatial data analysis and medical imaging [10]. The core problems and solutions in these domains differ significantly from the adversarial nature and real-time response requirements of cybersecurity, making them less directly relevant to the focus of this article.

Recent advancements in secure XAI frameworks for cybersecurity include the work presented in which introduces a conceptual framework for integrating XAI into Intrusion Detection Systems (IDS), providing practical guidelines for selecting XAI techniques based on operational requirements [11]. Another study reviews current trends and future opportunities in leveraging XAI to enhance IDS effectiveness, aiming for a more transparent and resilient cybersecurity landscape [12]. Furthermore, the researcher proposes MAESTRO, a threat modeling framework specifically designed for agentic AI, emphasizing the need for explainability and auditability in complex AI systems used in security [4]. The challenges and ethical considerations in AI-driven cybersecurity are discussed in another article, highlighting the ongoing need for research in trustworthy AI and robust privacy protection [13].

In the realm of AI trust calibration and human-AI collaboration, a multifaceted vision is presented in [14], emphasizing user empowerment and greater control in AI decision-making processes. The importance of a human-centered approach in designing AI solutions for seamless integration with existing processes is highlighted in [?], [15]. Research on adaptive trust calibration methods to improve the safety and efficiency of human-AI collaboration is explored in [16]. The critical role of human-AI partnership in enhancing cybersecurity, where AI augments human capabilities rather than replacing them, is discussed in [17].

Advancements in the realm of federated and adversarially robust models pertinent to cyber threat detection encompass the formulation of a secure federated learning framework tailored for malware triage, thereby enabling organizations to collaboratively train a global model while preserving the confidentiality of private data [18]. An innovative backdoor attack targeting federated learning, which strategically adapts to the evolving dynamics of the global model, is introduced in [19]. Investigations into the adversarial robustness of federated learning elucidate the detrimental effects of directly implementing adversarial training within federated contexts and propose a novel algorithm aimed at enhancing both accuracy and robustness [20]. Lastly, a secure detection framework that utilizes federated learning alongside robust aggregation protocols for the identification of advanced persistent threats within cyber-physical systems is delineated in [21].

IV. CONCLUSION

The proposed secure-XAI framework meticulously addresses the significant shortcomings in trust and explainability

that are characteristic of conventional black-box AI systems employed in the realm of cybersecurity. Through the integration of well-established explainability methodologies, resilient security infrastructures, and pivotal trust mechanisms, the framework aspires to furnish security professionals with a more transparent, dependable, and ultimately more efficacious assortment of AI-enhanced tools. This holistic approach equips analysts with the capability to comprehend the rationale underlying AI decisions, authenticate their precision, and retain authority over critical security operations.

However, the deployment of such a framework is not without potential challenges. A significant consideration is the inherent trade-off between the computational cost of generating explanations and the need for real-time performance in many cybersecurity applications. Balancing the desire for transparency with the practical constraints of operational environments will require careful optimization and potentially the selection of explainability methods that offer a suitable compromise. Furthermore, the seamless integration of diverse explainability, security, and trust techniques, along with ensuring effective human-AI collaboration, presents a complex engineering challenge.

Future work should focus on developing a functional prototype of the proposed framework to rigorously validate its feasibility and effectiveness in practical cybersecurity scenarios. Real-world evaluations in operational security environments are crucial to assess its impact on threat detection accuracy, analyst efficiency, and overall security posture. Additionally, ongoing research is needed to ensure that such secure-XAI frameworks align with evolving regulatory landscapes and ethical considerations in the field of cybersecurity, fostering responsible innovation and deployment of AI technologies.

REFERENCES

- [1] N. Dhir, H. Hoeltgebaum, N. M. Adams, M. Briers, A. Burke, and P. R. Jones, "Prospective artificial intelligence approaches for active cyber defence," *Cornell University*, 01 2021. [Online]. Available: <https://arxiv.org/abs/2104.09981>
- [2] S. Y. Liew, V. A. Hameed, M. E. Rana, and S. Safavi, "Navigating the retail 4.0 landscape: The transformative impact of cloud computing," 12 2023.
- [3] S. Safavi, Z. Shukur, M. E. Rana, U. A. Mokhtar, K. A. A. Bakar, N. b. Mohd Nazar, and F. I. Rumpa, "Securing the sustainable future of manufacturing: Analysis of iot retrofit challenges and solutions," 10 2024. [Online]. Available: <https://doi.org/10.20944/preprints202410.2045.v1>
- [4] S. Mane and D. J. Rao, "Explaining network intrusion detection system using explainable ai framework," *Cornell University*, 01 2021. [Online]. Available: <https://arxiv.org/abs/2103.07110>
- [5] Y. Bengio, G. E. Hinton, A. C.-C. Yao, D. Song, P. Abbeel, T. Darrell, Y. N. Harari, Y.-Q. Zhang, X. Lan, S. Shalev-Shwartz, G. K. Hadfield, J. Clune, T. Maharaj, F. Hutter, A. G. Baydin, S. A. McIlraith, Q. Gao, A. Acharya, D. Krueger, A. D. Dragan, P. H. S. Torr, S. Russell, D. Kahneman, J. Brauner, and S. Mindermann, "Managing extreme ai risks amid rapid progress," *American Association for the Advancement of Science*, vol. 384, no. 6698, pp. 842–845, 05 2024. [Online]. Available: <https://doi.org/10.1126/science.adn0117>
- [6] L. Alevizos and M. Dekker, "Towards an ai-enhanced cyber threat intelligence processing pipeline," *Multidisciplinary Digital Publishing Institute*, vol. 13, no. 11, pp. 2021–2021, 05 2024. [Online]. Available: <https://doi.org/10.3390/electronics13112021>
- [7] A. Tursunaliyeva, D. Alexander, R. Dunne, J. Li, L. G. Riera, and Y. Zhao, "Making sense of machine learning: A review of interpretation techniques and their applications," pp. 496–496, 01 2024. [Online]. Available: <https://doi.org/10.3390/app14020496>
- [8] S. Safavi and Z. Shukur, "Centeryou: A cloud-based approach to simplify android privacy management," 01 2020. [Online]. Available: <https://arxiv.org/abs/2008.13405>
- [9] S. Safavi, S. Safavi, and Z. Shukur, "Investigating the role of blockchain in enhancing home security: A comprehensive study," 12 2023, p. 1. [Online]. Available: <https://doi.org/10.1109/scored60679.2023.10563220>
- [10] S. Safavi, "Leveraging gis for environmental planning," vol. 22, p. 308.
- [11] F. Charnet, H. C. Tanuwidjaja, S. Ayoubi, P.-F. Gimenez, Y. Han, H. Jmila, G. Blanc, T. Takahashi, and Z. Zhang, "Explainable artificial intelligence for cybersecurity: a literature survey," *Springer Nature*, vol. 77, pp. 789–812, 10 2022. [Online]. Available: <https://doi.org/10.1007/s12243-022-00926-7>
- [12] G. Rjoub, J. Bentahar, O. A. Wahab, R. Mizouni, A. Song, R. Cohen, H. Otrok, and A. Mourad, "A survey on explainable artificial intelligence for cybersecurity," *Institute of Electrical and Electronics Engineers*, vol. 20, no. 4, pp. 5115–5140, 06 2023. [Online]. Available: <https://doi.org/10.1109/tnsm.2023.3282740>
- [13] K. Achuthan, S. Ramanathan, S. Srinivas, and R. Raman, "Advancing cybersecurity and privacy with artificial intelligence: current trends and future research directions," 12 2024. [Online]. Available: <https://doi.org/10.3389/fdata.2024.1497535>
- [14] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy ai: From principles to practices," pp. 1–46, 08 2022. [Online]. Available: <https://doi.org/10.1145/3555803>
- [15] M. Steyvers and A. Kumar, "Three challenges for ai-assisted decision-making," *SAGE Publishing*, vol. 19, no. 5, pp. 722–734, 07 2023. [Online]. Available: <https://doi.org/10.1177/17456916231181102>
- [16] K. Okamura and S. Yamada, "Adaptive trust calibration for human-ai collaboration," *Public Library of Science*, vol. 15, no. 2, 02 2020. [Online]. Available: <https://doi.org/10.1371/journal.pone.0229132>
- [17] N. Ezer, S. Bruni, Y. Cai, S. Hepenstal, C. Miller, and D. Schmorrow, "Trust engineering for human-ai teams," *SAGE Publishing*, vol. 63, no. 1, pp. 322–326, 11 2019. [Online]. Available: <https://doi.org/10.1177/1071181319631264>
- [18] A. Sadilek, L. Liu, D. T. Nguyen, M. Kamruzzaman, S. Serghiou, B. Rader, A. Ingerman, S. Mellem, P. Kairouz, E. O. Nsoesie, J. MacFarlane, A. Vullikanti, M. Marathe, P. R. Eastham, J. S. Brownstein, B. A. y Arcas, M. Howell, and J. Hernandez, "Privacy-first health research with federated learning," *Nature Portfolio*, vol. 4, no. 1, 09 2021. [Online]. Available: <https://doi.org/10.1038/s41746-021-00489-2>
- [19] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, and Y. Zhou, "A hybrid approach to privacy-preserving federated learning," pp. 1–11, 11 2019. [Online]. Available: <https://doi.org/10.1145/3338501.3357370>
- [20] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and P. S. Yu, "Privacy and robustness in federated learning: Attacks and defenses," *Institute of Electrical and Electronics Engineers*, vol. 35, no. 7, pp. 8726–8746, 11 2022. [Online]. Available: <https://doi.org/10.1109/tnnls.2022.3216981>
- [21] D. Preuveneers, V. Rimmer, I. Tsingenopoulos, J. Spooren, W. Joosen, and E. Ilie-Zudor, "Chained anomaly detection models for federated learning: An intrusion detection case study," *Multidisciplinary Digital Publishing Institute*, vol. 8, no. 12, pp. 2663–2663, 12 2018. [Online]. Available: <https://doi.org/10.3390/app8122663>